TI2V-Zero: Zero-Shot Image Conditioning for Text-to-Video Diffusion Models

MITSUBISHI ELECTRIC Changes for the Better | MITSUBISHI ELECTRIC RESEARCH LABORATORIES, INC | PennState | FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

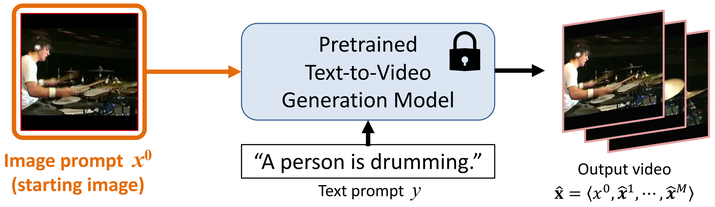
Haomiao Ni[1] Bernhard Egger[2] Suhas Lohit[3] Anoop Cherian[3] Ye Wang[3] Toshiaki Koike-Akino[3] Sharon X. Huang[1] Tim K. Marks[3]

[1]The Pennsylvania State University, USA [2]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany [3]Mitsubishi Electric Research Laboratories (MERL), USA

CVPR JUNE 17-21, 2024 SEATTLE, WA

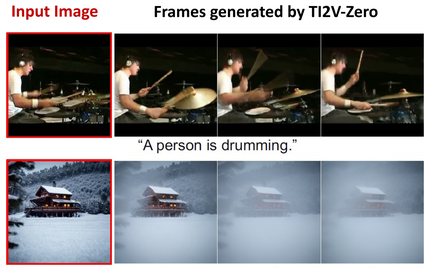
Motivation & Introduction

- **Text-conditioned Image-to-Video generation (TI2V)** aims to synthesize a realistic video from a given image (e.g., a photo of a person with drums) and a text prompt (e.g., "A person is drumming").
- Existing TI2V frameworks often require costly training on video-text datasets and specific model designs for text and image conditioning.
- Our proposed **TI2V-Zero** empowers a pretrained text-to-video (T2V) diffusion model to be conditioned on a provided image, enabling TI2V generation with **no optimization, fine-tuning, or external modules**.



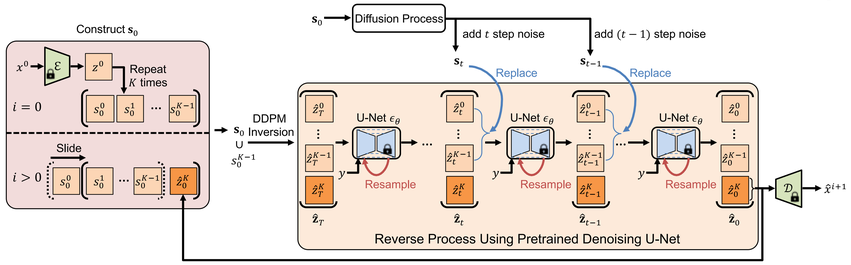
- **Contributions:**
  - A **"repeat-and-slide" strategy** that modulates the reverse denoising process of the frozen diffusion model to synthesize a video frame-by-frame starting from the provided image.
  - A **DDPM inversion strategy** to initialize Gaussian noise for each newly synthesized frame.
  - A **resampling technique** to help preserve visual details.

Example Results

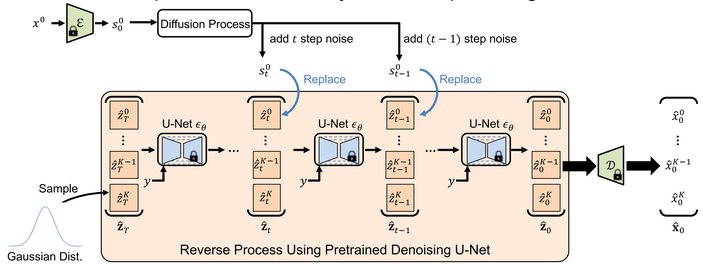


Methodology

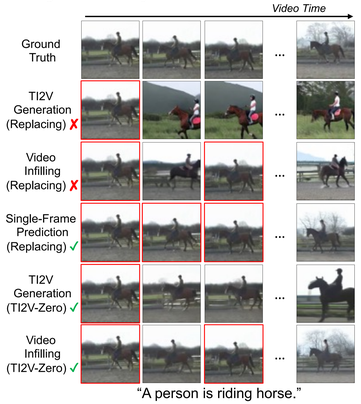
- **TI2V-Zero** forces temporal attention to use frames derived from the input image.



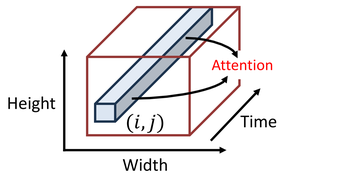
- **Our first idea:** Replace first frame by noised input image at each diffusion step.



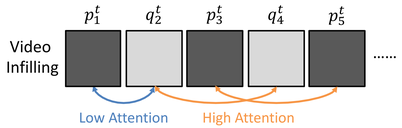
- **Analysis:** Why did our first idea fail?



Temporal attention layers in 3D denoising U-Net

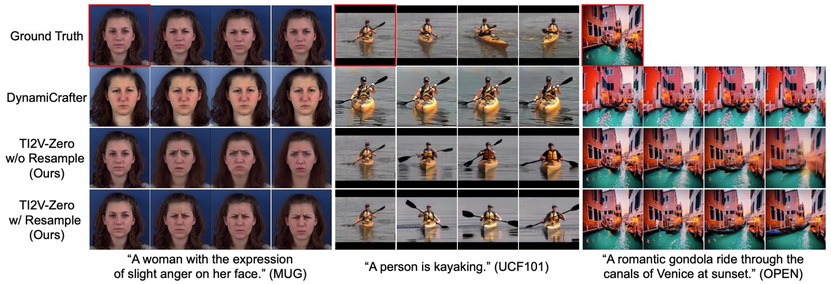


p^t: frames from adding noise to provided image.
q^t: frames from reducing noise of Gaussian sample.



Experiments

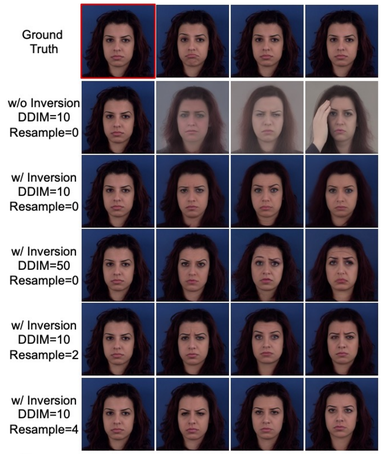
- **Datasets:** MUG (4 expressions), UCF101 (10 action classes), OPEN (10 text prompts from ChatGPT; 1,000 images from Stable Diffusion).



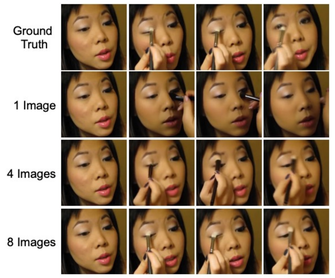
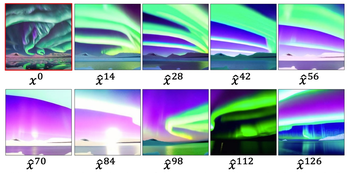
- **Metrics:** FVD, sFVD (subject-conditioned FVD), tFVD (text-conditioned FVD).

Model	MUG			UCF101	
	FVD↓	sFVD↓	tFVD↓	FVD↓	tFVD↓
DynamiCrafter [67]	1094.72	1359.86±257.73	1223.89±105.94	589.59	1540.02±199.59
TI2V-Zero w/o Resample (Ours)	339.89	443.97±139.10	405.22±61.58	493.19	1319.77±283.87
TI2V-Zero w/ Resample (Ours)	**180.09**	**267.17±74.72**	**252.77±39.02**	**477.19**	**1306.75±271.82**

- Ablation study comparing various sampling strategies:



- Video prediction from multiple frames:



- Long video generation (128 frames)